Inference-time Scaling for Diffusion-based Audio Super-resolution

Yizhu Jin[1], Zhen Ye[1], Zeyue Tian[1], Haohe Liu[2], Qiuqiang Kong[3], Yike Guo[1*], Wei Xue[1*]

[1]Hong Kong University of Science and Technology
[2]Meta AI
[3]Chinese University of Hong Kong
yikeguo@ust.hk, weixue@ust.hk

Abstract

Diffusion models have demonstrated remarkable success in generative tasks, including audio super-resolution (SR). In many applications like movie post-production and album mastering, substantial computational budgets are available for achieving superior audio quality. However, while existing diffusion approaches typically increase sampling steps to improve quality, the performance remains fundamentally limited by the stochastic nature of the sampling process, leading to high-variance and quality-limited outputs. Here, rather than simply increasing the number of sampling steps, we propose a different paradigm through inference-time scaling for SR, which explores multiple solution trajectories during the sampling process. Different task-specific verifiers are developed, and two search algorithms, including the random search and zero-order search for SR, are introduced. By actively guiding the exploration of the high-dimensional solution space through verifier-algorithm combinations, we enable more robust and higher-quality outputs. Through extensive validation across diverse audio domains (speech, music, sound effects) and frequency ranges, we demonstrate consistent performance gains, achieving improvements of up to 9.70% in aesthetics, 5.88% in speaker similarity, 15.20% in word error rate, and 46.98% in spectral distance for speech SR from 4 kHz to 24 kHz, showcasing the effectiveness of our approach.

Demo Page — https://racerk.github.io/tt-scale-audiosr

Introduction

Audio super-resolution (SR) aims to estimate high-frequency components from a low-resolution (LR) audio signal, thereby expanding its bandwidth and enhancing perceptual quality. Generally, audio SR is an ill-posed task, as the missing high-frequency content cannot be uniquely inferred from the observed low-frequency signal. In practice, this manifests as a one-to-many mapping problem: a single LR input may correspond to multiple plausible high-resolution (HR) outputs (Liu et al. 2021; Ye et al. 2023; Yu et al. 2023; Lee and Heo 2024). Deterministic models, which produce a single output that typically regresses to the average of all plausible HR outputs (Wang et al. 2018; Lee and Heo 2024), often fail to capture the mapping dynamics, limiting their ability to generate consistently high-quality HR predictions.

Recent advancements in diffusion models have shown great promise in modeling the distribution of high-dimensional data such as audio waveforms and spectrograms (Chen et al. 2024; Liu et al. 2023; Shen et al. 2023; Ye et al. 2024, 2023). Unlike traditional discriminative models, diffusion models learn to reverse a forward noise process, sampling from a Gaussian noise and iteratively transforming it into realistic data sample. Early works like NU-Wave (Lee and Han 2021) pioneers the use of diffusion model for audio SR, while NU-Wave2 (Han and Lee 2022) improves upon this by incorporating short-time Fourier convolution for more effective spectral modeling. To better handle mismatches between training and testing bandwidths—common in real-world scenarios where audio may be captured from diverse environments or devices with varying frequency responses, compression artifacts, or sampling rates, several models including NVSR (Liu et al. 2022), NU-Wave2 (Han and Lee 2022), and AudioSR (Liu et al. 2024) support flexible input bandwidths. Among them, AudioSR is notable for generalizing beyond the speech domain to more diverse audio content such as music and sound effects.

However, existing approaches have largely overlooked the role of inference-time randomness and uncertainty (Han and Lee 2022; Lee and Han 2021; Liu et al. 2024). Diffusion models inherently introduce sampling stochasticity, generating different HR outputs from the same LR input. Although this randomness is fundamental to the generative process (Ho, Jain, and Abbeel 2020; Song et al. 2020; Lu et al. 2022), its impact on perceptual audio quality has remained largely uncontrolled. In practice, we observe that perceptually similar LR signals can correspond to semantically distinct HR waveforms. For instance, in speech, different timbres or phonemes may be downsampled into nearly identical LR representations. Reversely, naive SR models often generate HR outputs with mismatched generation targets, which degrade intelligibility or alter speaker characteristics—evidenced by increased *Word Error Rate* and lower *Speaker Similarity* scores. Similarly, in music and sound effects, low-frequency components alone may not convey the full semantic content of the original signal, resulting in HR predictions that diverge from the intended meaning. These degradations in task-specific attributes call for targeted strategies during diffusion sampling process to restore lost attributes without retraining the model to retain both diversity and accuracy.

Recently, the study of Large Language Models has shown that allocating more computational resources at inference time through sophisticated search strategies can yield higher-quality and more contextually appropriate responses. This concept, known as *inference-time scaling*, highlights a promising direction for enhancing model performance without altering the training process (Brown et al. 2024; Snell et al. 2024; Ye et al. 2025). Analogously, inference-time scaling has been explored in diffusion models for vision tasks, where increasing the compute budget beyond denoising steps, has been shown to improve generative quality (Ma et al. 2025; Xie et al. 2025; Zhang et al. 2025c,d).

Inference-time scaling techniques remain largely underexplored in the audio domain for diffusion models (Zang, Li, and Kong 2025). In this work, we introduce a unified framework that applies inference-time scaling to improve audio SR quality by increasing the compute budget during inference. Specifically, our method samples multiple HR candidates using inference-time search algorithms and evaluates them using task-specific search verifiers, and selects the best-performing outputs. This approach allows us to navigate the solution space more effectively and recover critical perceptual qualities lost during vanilla SR generation. During this process, we find that over-optimizing with a single verifier can lead to overfitting and unintended artifacts (Clark et al. 2023; Pan, Bhatia, and Steinhardt 2022). To mitigate this, we ensemble multiple verifiers with complementary goals, enabling better trade-offs and more reliable improvements across diverse metrics.

Beyond performance gains, we quantify the range of the search space induced by different algorithms and examine the sample-wise variability of the diffusion process via uncertainty estimation.

Our key contributions are summarized as follows:

- We present the first systematic study of inference-time scaling for diffusion models in the audio domain, introducing a general framework that combines verifier-guided search with scalable compute algorithms to enhance perceptual quality across diverse audio types for the task of audio SR.
- We analyze verifier hacking effect and employ a verifier ensembling strategy to mitigate it, enabling better trade-offs across evaluation metrics. Our analysis further reveals that different task and upsampling settings exhibit distinct preferences for specific verifier–algorithm configurations.
- We quantitatively characterize the range of search space induced by different search algorithms, and perform uncertainty estimation of individual samples to reveal stochastic dynamics in the diffusion process for audio SR.

Related Work

Denoising Diffusion Probabilistic Models (DDPMs) have emerged as a leading class of generative models capable of producing high-fidelity outputs across a variety of domains, including images, 3D, audio, and video (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020; Rombach et al. 2022; Guo et al. 2023; Liu et al. 2023; Shen et al. 2023; García et al. 2025; Tian et al. 2025; Wu et al. 2025; Zhang et al. 2025b). DDPMs define a forward diffusion process that gradually corrupts clean data by adding Gaussian noise over T steps, and learn to reverse this process via a parameterized denoising network.

The original DDPM sampling process is computationally expensive due to its many iterative steps. To address this problem, Denoising Diffusion Implicit Model (DDIM) (Song, Meng, and Ermon 2020) introduces a deterministic, non-Markovian alternative to reverse sampling, enabling much faster generation while maintaining quality. Furthermore, to enable conditional generation without the need for explicit supervision, classifier-free guidance (Ho and Salimans 2022) allows the model to trade off between fidelity and conditioning strength by interpolating between unconditional and conditional predictions during inference.

As diffusion models scale to more complex data, such as high-resolution audio waveforms, efficiency becomes even more critical. Latent Diffusion Models (LDMs) (Rombach et al. 2022) address this challenge by moving the generative process into a learned latent space. A variational autoencoder first compresses the data into a lower-dimensional representation where diffusion is more efficient, and a decoder then reconstructs the signal back into its original form. In the audio domain, this approach proves especially effective due to the high temporal resolution and redundancy of waveform data. Instead of modeling raw waveforms directly, recent methods generate high-resolution mel-spectrograms in latent space, which are then converted into waveforms using neural vocoders (Ren et al. 2020, 2019; Ye et al. 2023). This two-stage architecture achieves both computational efficiency and high perceptual quality.

Methodology

Overview

We propose a unified inference-time scaling framework for audio SR with diffusion models, as illustrated in Figure 1. Our approach systematically explores the generative search space at inference by generating multiple HR audio candidates from an LR input and selecting the most promising output according to task-specific criteria. This is achieved by integrating two key components: (1) *search verifiers*, which evaluate the perceptual or semantic quality of each candidate, and (2) *search algorithms*, which efficiently traverse the candidate space guided by verifier feedback.

For the underlying audio SR model, we adopt the state-of-the-art AudioSR (Liu et al. 2024), which is based on LDM tailored for audio. AudioSR first predicts the HR Mel spectrogram conditioned on the LR input, and then reconstructs the waveform using a pretrained HiFiGAN-based vocoder (Kong, Kim, and Bae 2020; You et al. 2021). AudioSR supports a wide range of cutoff frequencies and audio types.

Following (Ma et al. 2025), our inference-time scaling framework is structured along two principal axes:

Verifiers are pretrained evaluation modules that assign scalar scores to generated HR audio, reflecting their quality with respect to specific tasks or conditions. Formally, a verifier is defined as

$$\mathcal{V}: \boldsymbol{X}^T \times \boldsymbol{C}^d \to \mathbb{R}, \tag{1}$$

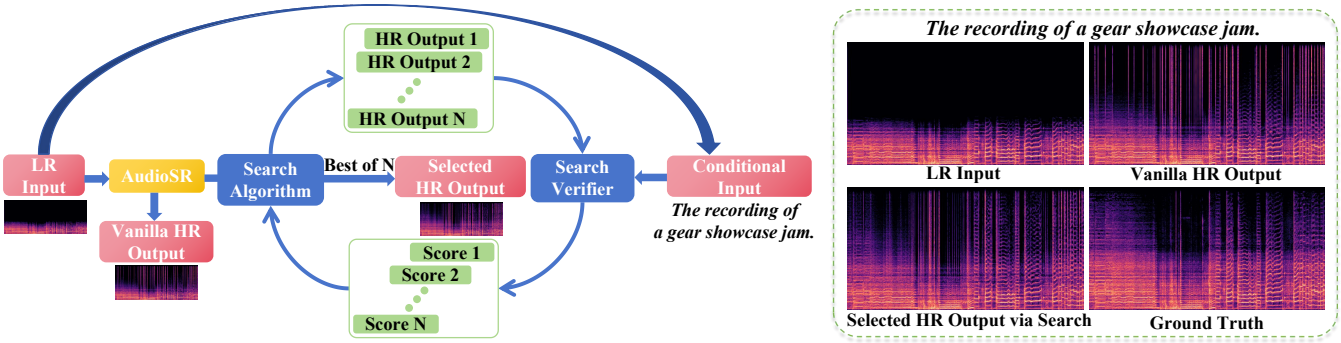


Figure 1: Overview of our inference-time scaling framework for audio SR. Given a LR input, multiple HR candidates are generated via diffusion sampling. A search algorithm explores this candidate space, guided by a verifier that scores each output based on a reference or task-specific criterion. The bottom row presents detailed STFT spectrograms of a music example. The selected output better aligns with the conditional text description and exhibits structural patterns closer to the reference.

where $\boldsymbol{X}^T$ denotes the HR waveform of length T, and $\boldsymbol{C}^d$ is an optional conditioning input (e.g., text, transcript, or audio prompt). Verifiers guide the ranking and selection of candidates, steering generation toward outputs that better satisfy task objectives.

Search Algorithms leverage verifier scores to identify the best HR candidate from a set of generated samples. Formally, a search algorithm is represented as

$$f: \mathcal{V} \times \mathcal{D}_\theta \times \{\boldsymbol{X}^T\}^N \times \boldsymbol{C}^d \rightarrow \boldsymbol{X}^T, \tag{2}$$

where $\mathcal{D}_\theta$ is the pretrained diffusion-based AudioSR model, and N is the number of candidate HR waveforms generated per inference. The algorithm selects the candidate with the highest verifier score, enabling systematic comparison across different verifier–algorithm configurations. For fair evaluation and efficient searching, we fix N during inference-time search for each method.

We further detail the search verifiers and search algorithms designed for audio SR as below.

Search Verifiers

We categorize search verifiers into two classes: *Oracle Verifier* and *Supervised Verifier*, based on their access to privileged information (Ma et al. 2025).

Oracle Verifier assumes access to ground-truth reference signals and directly evaluates each candidate using a full-reference metric. In the context of audio SR, we adopt the **Log-Spectrogram Distance (LSD) Verifier** as the oracle verifier. LSD computes the L2 distance between the log-magnitude Short-Time Fourier Transform (STFT) spectrograms of the generated and reference audio. By operating in the log-spectral domain, it emphasizes perceptually salient differences. LSD is widely used in speech and audio restoration tasks as a perceptual proxy for audio fidelity (Liu et al. 2024, 2022; Wang and Wang 2021).

Supervised Verifiers are employed under practical conditions where ground-truth reference signals are not accessible during inference. These verifiers consist of pretrained models that evaluate perceptual and semantic quality based on auxiliary conditioning inputs, rather than directly comparing against a reference signal.

For speech, we adopt the following verifiers:

- **Speaker Similarity (SpkSim) Verifier**: conditioned on a reference utterance from the target speaker, this verifier measures timbral consistency between the generated speech and the target speaker identity. It uses embeddings extracted from a pre-trained WavLM model (Chen et al. 2022).
- **Word Error Rate (WER) Verifier**: conditioned on target transcripts, it estimates transcription accuracy using a pre-trained automatic speech recognition (ASR) model (Gao et al. 2023; Radford et al. 2023) and computes the edit distance between predicted and reference text.
- **Aesthetics (AES) Verifier**: operating without any external conditioning, the AudioBox-Aesthetics model (Tjandra et al. 2025) offers no-reference audio quality assessment across four dimensions: Content Enjoyment (CE), Content Usefulness (CU), Production Complexity (PC), and Production Quality (PQ), providing a broad and interpretable evaluation for diverse audio types.

For non-speech audio such as music and sound effects, we utilize:

- **CLAP Verifier**: conditioned on textual descriptions, this verifier uses the Contrastive Language-Audio Pretraining (CLAP) model (Elizalde et al. 2023) to assess semantic alignment between the generated audio and its associated caption.
- **Aesthetics (AES) Verifier**: as described above, applied to non-speech audio for assessing overall perceptual quality in a modality-agnostic fashion.

To reflect real-world scenarios where ground-truth references are unavailable, we primarily use supervised verifiers to guide the search, while reserving the oracle verifier for evaluation purposes. To improve robustness and mitigate the issue of *verifier overfitting*—commonly referred to as *verifier hacking*, where the generation process may overly adapt to a specific verifier's scoring criteria, we further employ the

Ensemble Verifier. This approach combines the feedback of all relevant supervised verifiers for a given audio category. Due to differences in scoring scales across verifiers, we adopt a rank-based aggregation strategy: for each sample, we compute its relative rank under each component verifier, then average these ranks to obtain a unified ensemble score. This strategy is also applied internally within the Aesthetics Verifier, where we aggregate the rank scores across CE, CU, PC, and PQ to produce a holistic quality estimate.

Search Algorithms

Following the framework established in (Ma et al. 2025), we adopt two representative search algorithms: *Random Search* and *Zero-Order Search*.

Random Search (as shown in Algorithm 1) is the most straightforward strategy, implemented by sampling a set of N initial Gaussian noises from an isotropic distribution. Each noise sample is then passed through the DDIM sampler to generate HR outputs. The top-1 result is selected according to the verifier score. While simple, this method is prone to *verifier hacking*, as it explores the entire latent space without constraint, often exploiting verifier-specific biases (Clark et al. 2023; Ma et al. 2025; Pan, Bhatia, and Steinhardt 2022).

Zero-Order Search (as shown in Algorithm 2) incorporates iterative refinement around the selected pivot noise. This process starts with a randomly sampled noise $\mathbf{n}$, and then explores a neighborhood around it. Formally, a local neighborhood is defined as

$$S^{\lambda}_{\boldsymbol{n},i} = \{\mathbf{y} : d(\mathbf{y}, \mathbf{n}) = \lambda\}_{i=1}^{K}, \tag{3}$$

where $d(\cdot,\cdot)$ denotes a distance metric and λ defines the distance of the search. From this neighborhood, K candidates are generated and evaluated. The top-1 candidate is selected and then used as the new pivot noise. This process is repeated iteratively, gradually refining the search within a local region of the latent space.

Search Space Range Estimation

Existing research has yet to quantitatively characterize the *range* of the search space induced by different search verifier-algorithm combinations, primarily due to limitations in task granularity and evaluation precision. The audio SR task presents a particularly suitable setting to bridge this gap, as the generated HR outputs are structurally aligned with the LR inputs, yet exhibit substantial perceptual variability.

For a given algorithm-verifier pair, let $\mathcal{S}_N = \{x^{(i)}\}_{i=1}^{N}$ denote a set of N generated HR outputs. We define the *variance* of the search space as the average LSD between the STFT of each candidate and the mean spectrogram of the set:

$$\begin{cases} \mu_N = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \mathrm{STFT}(x^{(i)}) \\ \mathrm{Var}(\mathcal{S}_N) = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \mathrm{LSD}\left[\mathrm{STFT}(x^{(i)}), \mu_N\right] \end{cases} \tag{4}$$

This variance serves as a proxy for the search space's diversity, reflecting the spread of plausible HR estimations under the given inference-time configuration.

Algorithm 1: Random Search

Input: Pretrained DM $\mathcal{D}_\theta$,Verifier $\mathcal{V}$,Number of Candidates N
for $i = 1$ to N **do**
  Sample noise $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
  Generate sample $\boldsymbol{x}^{(i)} = \mathcal{D}_\theta(\boldsymbol{\epsilon}_i)$
  Evaluate score $\boldsymbol{s}^{(i)} = \mathcal{V}(\boldsymbol{x}^{(i)})$
end for
Select top-1 output $\boldsymbol{x}^* = \arg\max_i \boldsymbol{s}^{(i)}$
return $\boldsymbol{x}^*$

Algorithm 2: Zero-Order Search

Input: Initial noise $\boldsymbol{n}_0$, Search Distance λ, Neighbors K,
Pretrained DM $\mathcal{D}_\theta$, Verifier $\mathcal{V}$, Number of Candidates N
for $i = 1$ to N/K **do**
  for $k = 1$ to K **do**
    Sample noise $\boldsymbol{\epsilon}^{(i)}$ at distance λ
    Generate sample $\boldsymbol{x}^{(i)} = \mathcal{D}_\theta(\boldsymbol{\epsilon}^{(i)})$
    Evaluate score $\boldsymbol{s}^{(i)} = \mathcal{V}(\boldsymbol{x}^{(i)})$
  end for
  Update $\boldsymbol{n}_0 = \boldsymbol{\epsilon}^{(i^*)}, i^* = \arg\max_i \boldsymbol{s}^{(i)}$
end for
return $\mathcal{D}_\theta(\boldsymbol{n}_0)$

Uncertainty Estimation

Recent advances in image SR have highlighted the value of uncertainty modeling. For example, (Ning et al. 2021) integrates spatial uncertainty into the training loss to enforce stronger supervision in ambiguous regions. Similarly, (Zhang et al. 2025a) leverages deterministic models to estimate uncertainty via residuals between downsampled and upsampled images, guiding region-specific noise control for better reconstruction. However, such residual-based strategies are less applicable to audio SR due to the absence of a well-defined HR spectrogram after downsampling and upsampling operations. As a result, residuals fail to capture the true ambiguity of HR estimation.

To address this, we propose to estimate uncertainty directly from the stochastic nature of diffusion sampling. We compute an *uncertainty map* (Zhang et al. 2025a) by measuring the variance across time-frequency bins in the STFT domain over multiple generations from the same LR input. This approach reveals fine-grained regions of variability, providing insight into the ill-posedness of the task and the sensitivity of different spectral regions to sampling noise. Specifically, we estimate the variance at each time-frequency bin across all STFTs in $\mathcal{S}_N$, and normalize the resulting values linearly:

$$\begin{cases} \mathcal{U}(t,f) = \dfrac{\mathrm{Var}\left[\mathrm{STFT}(x^{(i)})_{t,f}\right]_{i=1}^{N} - \min}{\max - \min + \epsilon}, \\ \min = \min \mathrm{Var}\left[\mathrm{STFT}(x^{(i)})_{t,f}\right]_{i=1}^{N}, \\ \max = \max \mathrm{Var}\left[\mathrm{STFT}(x^{(i)})_{t,f}\right]_{i=1}^{N}, \end{cases} \tag{5}$$

where $\mathcal{U}(t,f)$ denotes the normalized uncertainty score at time frame t and frequency bin f, and ϵ is a small constant for

numerical stability. This formulation highlights regions with high generative variance, revealing time-frequency structures that are inherently ambiguous or sensitive to the generative stochasticity of diffusion models.

Experiments

Evaluation Datasets

We construct three curated evaluation benchmarks, each comprising 200 samples, to assess model performance across different audio types: VCTK (Liu et al. 2022) for speech, MusicCaps (Agostinelli et al. 2023) for music, and ESC-50 (Piczak 2015) for sound effects. These datasets are tailored to reflect the specific characteristics and evaluation criteria of each audio domain. The detailed settings for each audio type are as follows.

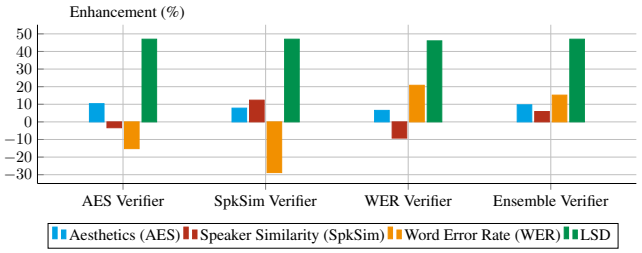
For *Speech*, to evaluate intelligibility and speaker consistency, we use the ground-truth transcripts provided by VCTK to compute the Word Error Rate (WER). Additionally, we randomly sample other utterances from the same speaker to serve as audio references for the Speaker Similarity (SpkSim) Verifier. For *Music*, to ensure evaluation quality, we first pre-filter low-quality samples using MusicCaps captions by discarding entries containing keywords such as *"mediocre"*, *"low quality"*, or *"low fidelity"*. From the remaining subset, we select samples that span diverse genres and exhibit rich high-frequency content, thereby better testing the SR capability. For *Sound Effect*, since ESC-50 lacks human-written descriptions, we generate audio captions using Qwen2-Audio (Chu et al. 2024), conditioned on the original ESC-50 category keywords. These synthesized captions are then paired with their corresponding audio clips and used in CLAP Verifier, which evaluates semantic alignment via contrastive audio-text embeddings.

Search Settings

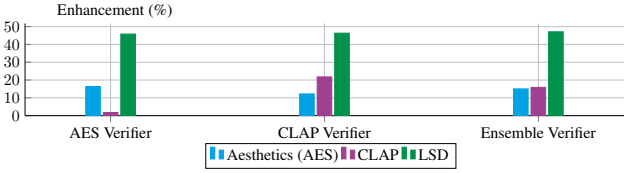
For AudioSR, we retain its default settings using 50 DDIM sampling steps and classifier-free guidance scale of 3.5 during inference-time sampling. We fix the size of the inference-time search space to $N = 120$ for each algorithm-verifier pair. For each configuration, we select the top-1 HR output according to the verifier scores. We denote *Random Search* and *Zero-Order Search* as **Random** and **Zero-Order**, respectively. The *Zero-Order Search* algorithm is configured with $K = 2$ neighborhood candidates and a search distance parameter of $\lambda = 0.99$.

Experiment Result Analysis

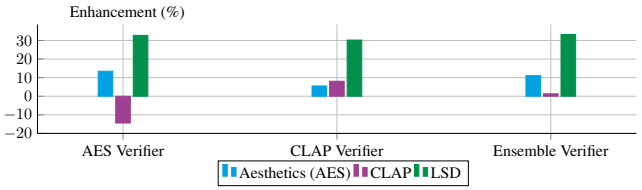
Effectiveness of Scaling Inference-Time Compute. As shown in Table 1 and Figure 2, we present a comprehensive evaluation of performance improvements across verifier-algorithm combinations for speech, music, and sound effects at cutoff frequencies of 4 kHz and 8 kHz. While vanilla audio SR improves perceptual fidelity in general, it often comes at the cost of degrading critical attributes when compared to the LR input. For speech, this degradation is particularly evident in metrics such as Speaker Similarity and WER, which reflect speaker consistency and intelligibility. Similarly, for music and sound effects, we observe a consistent drop in



(a) Speech, from 4 kHz to 24 kHz.



(b) Music, from 4 kHz to 24 kHz.



(c) Sound Effect, from 4 kHz to 24 kHz.

Figure 2: Performance improvements over the LR input across different audio types using inference-time *Random Search* with various verifiers from 4 kHz to 24 kHz. The *enhancement* denotes the relative improvement over LR under each evaluation metric. For Aesthetics (AES), Speaker Similarity (SpkSim) and CLAP Score, enhancements reflect relative increases. For Word Error Rate (WER) and Log Spectrogram Distance (LSD), improvements are computed as relative reductions.

CLAP Score, which measures the alignment between audio and semantic content. These findings underscore the limitations of vanilla SR generation and motivate the use of verifier-guided inference-time scaling strategies to restore such domain-specific qualities. Crucially, we find that when the verifier used for search directly aligns with the evaluation metric—such as using the Speaker Similarity Verifier to optimize Speaker Similarity, or the CLAP Verifier to improve CLAP Score—*Random Search* consistently outperforms *Zero-Order Search* in recovering the lost performance. This is largely due to its ability to explore a broader candidate space, enabling more effective correction of deficiencies introduced by the vanilla SR process.

Search Space Range Estimation. Empirical evidence suggests that *Random Search* has a higher likelihood of locating global optima, whereas *Zero-Order Search* exhibits stronger locality due to its iterative refinements around selected initial noise samples (Ma et al. 2025). To quantitatively assess this behavior, we estimate the search space range using Equation 4. As shown in Table 2, the search space range tends to be larger for lower cutoff frequencies compared to their

		Speech							
		4 kHz				8 kHz			
		AES(↑)	SpkSim(↑)	WER(↓)	LSD(↓)	AES(↑)	SpkSim(↑)	WER(↓)	LSD(↓)
LR Input		4.74	0.510	0.125	3.15	5.16	0.596	0.116	2.84
Vanilla AudioSR		4.74	*0.370*	*0.263*	1.73	5.18	*0.535*	*0.136*	1.62
AES Verifier	+ Random	**5.23**	*0.493*	*0.144*	1.67	**5.32**	*0.572*	*0.123*	**1.53**
	+ Zero-Order	5.03	*0.458*	*0.156*	**1.66**	5.23	*0.570*	0.116	**1.53**
SpkSim Verifier	+ Random	5.11	**0.573**	*0.161*	1.67	5.25	**0.617**	0.116	1.58
	+ Zero-Order	4.99	*0.471*	*0.165*	**1.66**	5.22	*0.582*	*0.121*	**1.53**
WER Verifier	+ Random	5.05	*0.463*	**0.099**	1.70	5.25	*0.575*	**0.105**	1.58
	+ Zero-Order	4.98	*0.458*	*0.154*	**1.66**	5.21	*0.575*	0.114	**1.53**
Ensemble Verifier	+ Random	5.20	0.540	0.106	1.67	5.31	0.601	0.106	1.55
	+ Zero-Order	5.02	*0.469*	*0.152*	**1.66**	5.23	*0.578*	0.113	**1.53**

		Music						Sound Effect					
		4 kHz			8 kHz			4 kHz			8 kHz		
		AES(↑)	CLAP(↑)	LSD(↓)	AES(↑)	CLAP(↑)	LSD(↓)	AES(↑)	CLAP(↑)	LSD(↓)	AES(↑)	CLAP(↑)	LSD(↓)
LR Input		6.06	0.340	3.95	6.47	0.352	3.09	4.16	0.458	3.97	4.42	0.481	3.33
Vanilla AudioSR		6.57	*0.303*	2.20	6.63	*0.323*	2.05	4.22	*0.343*	2.86	4.43	*0.421*	2.49
AES Verifier	+ Random	**7.05**	0.346	2.14	**6.94**	0.358	2.03	**4.72**	*0.392*	2.67	**4.72**	*0.455*	2.49
	+ Zero-Order	6.81	0.344	2.15	6.79	0.352	**2.02**	4.45	*0.398*	2.76	4.54	*0.457*	2.44
CLAP Verifier	+ Random	6.80	**0.414**	2.12	6.75	**0.424**	2.04	4.39	**0.495**	2.77	4.51	**0.522**	2.53
	+ Zero-Order	6.77	0.357	2.16	6.77	0.373	2.05	4.39	*0.409*	2.74	4.50	*0.471*	**2.43**
Ensemble Verifier	+ Random	6.97	0.394	**2.09**	6.89	0.404	2.03	4.62	0.464	**2.65**	4.65	0.502	2.48
	+ Zero-Order	6.79	0.352	2.18	6.78	0.368	2.03	4.41	*0.415*	2.72	4.54	*0.469*	**2.43**

Table 1: Comprehensive evaluation of inference-time scaling across verifier-algorithm combinations for speech, music, and sound effects at cutoff frequencies of 4 kHz and 8 kHz. **Bold** indicates top-1 performance, underline denotes top-2, and *italics* highlight cases where the performance of the generated output is worse than the LR input. AES, SpkSim, WER, and LSD stand for Aesthetics Score, Speaker Similarity, Word Error Rate, and Log-Spectrogram Distance, respectively. The Ensemble Verifier aggregates AES, SpkSim, WER scores for speech, and AES, CLAP scores for music and sound effects. Random and Zero-Order denote *Random Search* and *Zero-Order Search*, respectively.

Algorithm	4 kHz			8 kHz			**Average**
	Speech	Music	Sound Effect	Speech	Music	Sound Effect	**LSD Variance**
Random	**0.673**	**0.921**	**1.25**	**0.577**	**0.868**	**1.10**	**0.898**
Zero-Order	0.608	0.790	0.947	0.529	0.750	0.940	0.761

Table 2: Comparison of search range across algorithms based on LSD variance for different audio types at 4 kHz, 8 kHz, and their average. Random and Zero-Order denote *Random Search* and *Zero-Order Search*, respectively.

higher-frequency counterparts. Moreover, we observe a progressive increase in search space range from speech to music and then to sound effects, suggesting rising levels of uncertainty across these audio types. Consistently, *Random Search* exhibits a significantly wider search range than *Zero-Order Search*, confirming its advantage in exploring diverse generative candidates.

Uncertainty Estimation. One of the central challenges in audio SR with diffusion models stems from the inherent randomness of the denoising process, which leads to significant output variability across different sampling runs. However, this stochastic behavior is often under-characterized. To address this, we employ uncertainty map visualization as an interpretability tool to highlight the localized variance patterns embedded within individual samples, as shown in Figure 4. For implementation, we compute the variance across the time-frequency bins of STFT spectrograms derived from *Random Search* generation candidates, since this algorithm has a wider search range. To enhance visual interpretability, we apply percentile-based contrast normalization by clipping values above the 90th percentile, thus emphasizing the structure of salient uncertainty regions prior to rendering.

Verifier Hacking. To examine the implications of scaling up inference-time compute, we focus our analysis on *Random Search*, which offers a wider search range. This setting allows us to better isolate the phenomenon of *verifier hacking*, where the optimization overfits to a specific verifier without achieving meaningful or comprehensive improvements. As shown in Figure 3, we present the relative performance improvements while scaling up inference-time compute budget in *Random Search* over the default generation in speech from 4 kHz to 24 kHz. Among the verifiers, LSD Verifier

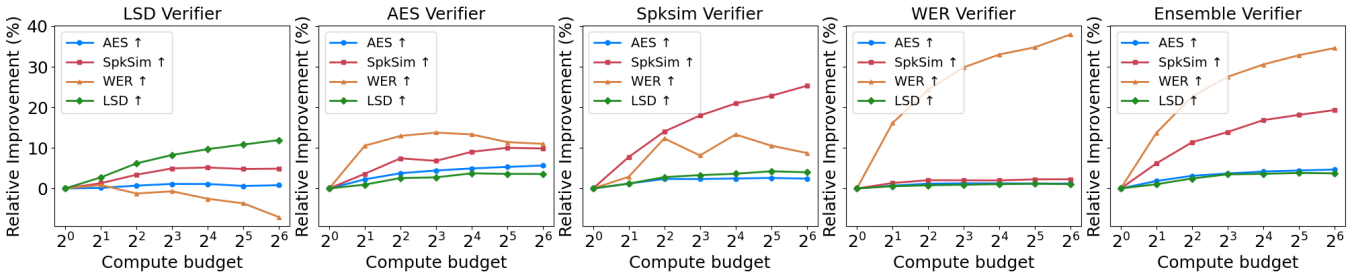


Figure 3: Relative performance improvements over the default generation (vanilla AudioSR) for speech from 4 kHz to 24 kHz in *Random Search*, demonstrating the effect of inference-time scaling across different verifier types. LSD, AES, SpkSim, and WER refer to Log Spectrogram Distance, Aesthetics Score, Speaker Similarity, and Word Error Rate, respectively. The Ensemble Verifier aggregates AES, SpkSim, and WER by averaging their rank scores.

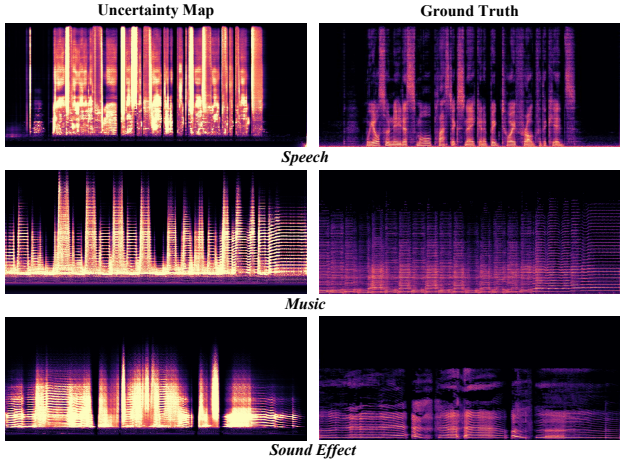


Figure 4: Visualization of uncertainty maps over STFT spectrograms across diverse audio types. Brighter regions indicate higher generative uncertainty across diffusion samples. Overlapping high-uncertainty areas reflect divergent spectral realizations, underscoring the necessity of inference-time scaling to identify perceptually optimal outputs.

serves as *Oracle Verifier*, with an access to ground-truth references while searching. Aesthetics (AES), Speaker Similarity (SpkSim), Word Error Rate (WER), CLAP and Ensemble Verifiers are treated as *Supervised Verifiers*, which we employ at practical settings without referring to ground-truth signals while searching.

Among the evaluation metrics for speech, WER behaves uniquely. Notably, when guided by the *Oracle Verifier*, LSD Verifier, we observe that higher LSD scores do not necessarily correspond to improvements in essential speech attributes, particularly WER. In fact, WER often worsens even as LSD improves, revealing a misalignment between general fidelity and intelligibility. This highlights the need for *Supervised Verifiers* during inference-time search to preserve perceptual quality specific to the speech SR task.

Furthermore, while the WER Verifier is effective in improving WER alone, it provides limited gains in other metrics, emphasizing its narrow focus. As we scale up the inference-time compute budget, we observe that both the Aesthetics and Speaker Similarity Verifiers begin to exhibit diminishing or even negative returns on WER beyond a search space size of approximately $2^3 \sim 2^4$, indicating a *verifier hacking* phenomenon, where the search process overfits to the verifier at the cost of overall quality. In contrast, the *Ensemble Verifier* mitigates such overfitting by averaging the ranks from multiple verifiers (AES, SpkSim, and WER), leading to a more balanced trade-off across all metrics and achieving meaningful improvements in WER without compromising other perceptual qualities.

Verifier-Algorithm-Task Alignment. When upsampling from 4 kHz to 24 kHz, the combination of *Ensemble Verifier* and *Random Search* consistently achieves the best overall trade-off across audio types. As the SR setting shifts from 4 kHz to 8 kHz inputs (both targeting 24 kHz), different audio domains exhibit distinct verifier–algorithm preferences. For speech, this combination remains superior due to its balanced enhancement of intelligibility and timbre, while for music and sound effects, we observe a stronger affinity toward the *Aesthetics* and *CLAP* verifiers, respectively.

Conclusion

This work presents a comprehensive study of inference-time scaling for diffusion models in the context of audio SR. We propose a unified framework that explores scalable search strategies guided by diverse verifier-algorithm combinations, achieving consistent improvements across audio types and cutoff frequencies. We further identify the phenomenon of verifier hacking and demonstrate that verifier ensembling effectively mitigates this issue by balancing competing perceptual objectives. Beyond performance gains, we quantify search space ranges via LSD variance, revealing how search dynamics vary across domains and upsampling settings. We also introduce variance-based uncertainty maps to highlight time-frequency regions sensitive to generative noise, offering deeper insights into the stochasticity and ill-posedness of the audio SR task, while opening up promising directions for future research in uncertainty-aware diffusion modeling in audio generation.

References

Agostinelli, A.; Denk, T. I.; Borsos, Z.; Engel, J.; Verzetti, M.; Caillon, A.; Huang, Q.; Jansen, A.; Roberts, A.; Tagliasacchi, M.; et al. 2023. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*.

Brown, B.; Juravsky, J.; Ehrlich, R.; Clark, R.; Le, Q. V.; Ré, C.; and Mirhoseini, A. 2024. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*.

Chen, K.; Wu, Y.; Liu, H.; Nezhurina, M.; Berg-Kirkpatrick, T.; and Dubnov, S. 2024. Musicldm: Enhancing novelty in text-to-music generation using beat-synchronous mixup strategies. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1206–1210. IEEE.

Chen, S.; Wang, C.; Chen, Z.; Wu, Y.; Liu, S.; Chen, Z.; Li, J.; Kanda, N.; Yoshioka, T.; Xiao, X.; et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6): 1505–1518.

Chu, Y.; Xu, J.; Yang, Q.; Wei, H.; Wei, X.; Guo, Z.; Leng, Y.; Lv, Y.; He, J.; Lin, J.; et al. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Clark, K.; Vicol, P.; Swersky, K.; and Fleet, D. J. 2023. Directly fine-tuning diffusion models on differentiable rewards. *arXiv preprint arXiv:2309.17400*.

Elizalde, B.; Deshmukh, S.; Al Ismail, M.; and Wang, H. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Gao, Z.; Li, Z.; Wang, J.; Luo, H.; Shi, X.; Chen, M.; Li, Y.; Zuo, L.; Du, Z.; Xiao, Z.; et al. 2023. Funasr: A fundamental end-to-end speech recognition toolkit. *arXiv preprint arXiv:2305.11013*.

García, H. F.; Nieto, O.; Salamon, J.; Pardo, B.; and Seetharaman, P. 2025. Sketch2sound: Controllable audio generation via time-varying signals and sonic imitations. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Guo, Y.; Yang, C.; Rao, A.; Liang, Z.; Wang, Y.; Qiao, Y.; Agrawala, M.; Lin, D.; and Dai, B. 2023. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*.

Han, S.; and Lee, J. 2022. NU-Wave 2: A general neural audio upsampling model for various sampling rates. *arXiv preprint arXiv:2206.08545*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Kong, J.; Kim, J.; and Bae, J. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33: 17022–17033.

Lee, J.; and Han, S. 2021. Nu-wave: A diffusion probabilistic model for neural audio upsampling. *arXiv preprint arXiv:2104.02321*.

Lee, M.; and Heo, J.-P. 2024. Noise-free optimization in early training steps for image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2920–2928.

Liu, H.; Chen, K.; Tian, Q.; Wang, W.; and Plumbley, M. D. 2024. AudioSR: Versatile audio super-resolution at scale. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1076–1080. IEEE.

Liu, H.; Chen, Z.; Yuan, Y.; Mei, X.; Liu, X.; Mandic, D.; Wang, W.; and Plumbley, M. D. 2023. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*.

Liu, H.; Choi, W.; Liu, X.; Kong, Q.; Tian, Q.; and Wang, D. 2022. Neural vocoder is all you need for speech super-resolution. *arXiv preprint arXiv:2203.14941*.

Liu, H.; Kong, Q.; Tian, Q.; Zhao, Y.; Wang, D.; Huang, C.; and Wang, Y. 2021. VoiceFixer: Toward general speech restoration with neural vocoder. *arXiv preprint arXiv:2109.13731*.

Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; and Zhu, J. 2022. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35: 5775–5787.

Ma, N.; Tong, S.; Jia, H.; Hu, H.; Su, Y.-C.; Zhang, M.; Yang, X.; Li, Y.; Jaakkola, T.; Jia, X.; et al. 2025. Scaling Inference Time Compute for Diffusion Models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2523–2534.

Ning, Q.; Dong, W.; Li, X.; Wu, J.; and Shi, G. 2021. Uncertainty-driven loss for single image super-resolution. *Advances in Neural Information Processing Systems*, 34: 16398–16409.

Pan, A.; Bhatia, K.; and Steinhardt, J. 2022. The effects of reward misspecification: Mapping and mitigating misaligned models. *arXiv preprint arXiv:2201.03544*.

Piczak, K. J. 2015. ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, 1015–1018.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, 28492–28518. PMLR.

Ren, Y.; Hu, C.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; and Liu, T.-Y. 2020. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*.

Ren, Y.; Ruan, Y.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; and Liu, T.-Y. 2019. Fastspeech: Fast, robust and controllable text to speech. *Advances in neural information processing systems*, 32.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Shen, K.; Ju, Z.; Tan, X.; Liu, Y.; Leng, Y.; He, L.; Qin, T.; Zhao, S.; and Bian, J. 2023. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. *arXiv preprint arXiv:2304.09116*.

Snell, C.; Lee, J.; Xu, K.; and Kumar, A. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Tian, Z.; Jin, Y.; Liu, Z.; Yuan, R.; Tan, X.; Chen, Q.; Xue, W.; and Guo, Y. 2025. Audiox: Diffusion transformer for anything-to-audio generation. *arXiv preprint arXiv:2503.10522*.

Tjandra, A.; Wu, Y.-C.; Guo, B.; Hoffman, J.; Ellis, B.; Vyas, A.; Shi, B.; Chen, S.; Le, M.; Zacharov, N.; et al. 2025. Meta Audiobox Aesthetics: Unified Automatic Quality Assessment for Speech, Music, and Sound. *arXiv preprint arXiv:2502.05139*.

Wang, H.; and Wang, D. 2021. Towards robust speech super-resolution. *IEEE/ACM transactions on audio, speech, and language processing*, 29: 2058–2066.

Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; and Change Loy, C. 2018. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, 0–0.

Wu, J. Z.; Zhang, Y.; Turki, H.; Ren, X.; Gao, J.; Shou, M. Z.; Fidler, S.; Gojcic, Z.; and Ling, H. 2025. Difix3d+: Improving 3d reconstructions with single-step diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 26024–26035.

Xie, E.; Chen, J.; Zhao, Y.; Yu, J.; Zhu, L.; Wu, C.; Lin, Y.; Zhang, Z.; Li, M.; Chen, J.; et al. 2025. Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer. *arXiv preprint arXiv:2501.18427*.

Ye, Z.; Ju, Z.; Liu, H.; Tan, X.; Chen, J.; Lu, Y.; Sun, P.; Pan, J.; Bian, W.; He, S.; et al. 2024. Flashspeech: Efficient zero-shot speech synthesis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 6998–7007.

Ye, Z.; Xue, W.; Tan, X.; Chen, J.; Liu, Q.; and Guo, Y. 2023. Comospeech: One-step speech and singing voice synthesis via consistency model. In *Proceedings of the 31st ACM International Conference on Multimedia*, 1831–1839.

Ye, Z.; Zhu, X.; Chan, C.-M.; Wang, X.; Tan, X.; Lei, J.; Peng, Y.; Liu, H.; Jin, Y.; DAI, Z.; et al. 2025. Llasa: Scaling Train-Time and Inference-Time Compute for Llama-based Speech Synthesis. *arXiv preprint arXiv:2502.04128*.

You, J.; Kim, D.; Nam, G.; Hwang, G.; and Chae, G. 2021. GAN vocoder: Multi-resolution discriminator is all you need. *arXiv preprint arXiv:2103.05236*.

Yu, C.-Y.; Yeh, S.-L.; Fazekas, G.; and Tang, H. 2023. Conditioning and sampling in variational diffusion models for speech super-resolution. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Zang, Y.; Li, J.; and Kong, Q. 2025. Training-Free Multi-Step Audio Source Separation. *arXiv preprint arXiv:2505.19534*.

Zhang, L.; You, W.; Shi, K.; and Gu, S. 2025a. Uncertainty-guided Perturbation for Image Super-Resolution Diffusion Model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 17980–17989.

Zhang, S.; Li, W.; Chen, S.; Ge, C.; Sun, P.; Zhang, Y.; Jiang, Y.; Yuan, Z.; Peng, B.; and Luo, P. 2025b. FlashVideo: Flowing Fidelity to Detail for Efficient High-Resolution Video Generation. *arXiv preprint arXiv:2502.05179*.

Zhang, X.; Lin, H.; Ye, H.; Zou, J.; Ma, J.; Liang, Y.; and Du, Y. 2025c. Inference-time Scaling of Diffusion Models through Classical Search. *arXiv preprint arXiv:2505.23614*.

Zhang, Z.; Xie, J.; Lu, Y.; Yang, Z.; and Yang, Y. 2025d. In-context edit: Enabling instructional image editing with in-context generation in large scale diffusion transformer. *arXiv preprint arXiv:2504.20690*.